Proceedings of the 2013 Winter Simulation Conference
R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, eds.

HAVE WE REALLY BEEN ANALYZING TERMINATING SIMULATIONS INCORRECTLY ALL THESE YEARS?

Paul J. Sánchez

Operations Research
Naval Postgraduate School
1411 Cunningham Road
Monterey, CA 93943, USA

K. Preston White, Jr.

Department of Systems Engineering
University of Virginia
PO Box 400747
Charlottesville, VA 22976, USA

ABSTRACT

We all know how to estimate a confidence interval for the mean based on a random sample. The interval is centered on the sample mean, with the half-width proportional to the sample standard error. We know also that terminating simulations generate independent observations. What simulators appear to have overlooked is that independence alone is insufficient to guarantee a valid random sample—the observations must also be identically distributed. This is a good assumption if the outcome of each replication is a single observation, but it is demonstrably incorrect if the outcome is an aggregate value and the replications have differing numbers of observations. In this paper we explore the implications of this oversight when within-replication observations are independent. We then derive analytic results showing that although the impact on interval estimates can sometimes be negligible, there also are circumstances where the variance of our estimates is significantly increased. We finish with a simple example which demonstrates the potential impact for practitioners.

1 INTRODUCTION

Analysis of *terminating simulations*, i.e., models that halt when they reach some clearly defined state, seems straightforward (Banks et al. 2000; Bratley, Fox, and Schrage 1983; Hoover and Perry 1989; Law and Kelton 2000). If each run is seeded independently of the others, then the output measures from each run will be independent and we can just apply classical statistics, right? It turns out that the answer to that question may be "wrong!" The "identically distributed" requirement of classical statistics can be called into question when the output measure is an aggregate such as a sum or sample mean.

Performance measures of a simulation will always be some function of the simulation state. In the case of terminating simulations, the performance measure is reported upon termination. Each replication will produce one observation of the performance measure. If that observation directly represents an end state such as the number of failed components after a week's operation, or the number of patients processed in 24 hours of emergency room operations, there's no problem—the set of values obtained by replication represent a random sample from the distribution of possible end states, and classical statistics applies. However, we can run into trouble if the performance measure is an aggregate measure, such as an average, and the number of observations contributing to the aggregate varies from replication to replication.

2 BACKGROUND AND NOTATION

Let Y_{ij} be the j^{th} raw observation from the i^{th} replication of our model, with common mean μ_Y and variance σ_Y^2. For this paper we will assume that we are performing r independently seeded replications of our model,

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
DEC 2013		**00-00-2013 to 00-00-2013**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Have We Really Been Analyzing Terminating Simulations Incorrectly All These Years?	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Naval Postgraduate School,Operations Research Department,Monterey,CA,93943**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Proceedings of the 2013 Winter Simulation Conference, 8-11 Dec, Washington, DC.

14. ABSTRACT
We all know how to estimate a confidence interval for the mean based on a random sample. The interval is centered on the sample mean, with the half-width proportional to the sample standard error. We know also that terminating simulations generate independent observations. What simulators appear to have overlooked is that independence alone is insufficient to guarantee a valid random sample???the observations must also be identically distributed. This is a good assumption if the outcome of each replication is a single observation but it is demonstrably incorrect if the outcome is an aggregate value and the replications have differing numbers of observations. In this paper we explore the implications of this oversight when within-replication observations are independent. We then derive analytic results showing that although the impact on interval estimates can sometimes be negligible, there also are circumstances where the variance of our estimates is significantly increased. We finish with a simple example which demonstrates the potential impact for practitioners.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**7**	

the i^{th} replication of our terminating simulation produces a single aggregate measure

$$\bar{Y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$$

for $i = 1 \ldots r$ based on n_i observations. Note that n_i is explicitly allowed to vary from replication to replication. We will denote the total number of observations as $N = \sum_{i=1}^{r} n_i$. The Y_{ij}'s are independent across i by virtue of independent seeding, and in this paper we will assume that they are also independent within a replication to emphasize the effect of varying the sample sizes. Given these assumptions, it should be clear that $E[\bar{Y}_i] = \mu_Y$ and $Var(\bar{Y}_i) = \sigma_Y^2/n_i$. They are *not* identically distributed, because the variance varies from run to run.

We needn't worry about initial bias in a terminating scenario. The traditional wisdom at this point would be to calculate a grand sample mean and estimate the variance across the replications:

$$\bar{\bar{Y}} = \frac{1}{r}\sum_{i=1}^{r} \bar{Y}_i \tag{1}$$

$$s^2_{\bar{\bar{Y}}} = \frac{1}{r-1}\sum_{i=1}^{r} \left(\bar{Y}_i - \bar{\bar{Y}}\right)^2. \tag{2}$$

and use them to construct a $100(1-\alpha)\%$ confidence interval of the form

$$\bar{\bar{Y}} \pm t_{\alpha/2;r-1}\sqrt{\frac{s^2_{\bar{\bar{Y}}}}{r}}$$

where $t_{\alpha/2;r-1}$ is the critical value from Student's T distribution with $r-1$ degrees of freedom and probability $\alpha/2$ in each tail.

3 IMPACT OF UNEQUAL SAMPLE SIZES

So what is the impact on $\bar{\bar{Y}}$ of using $\bar{Y}_i$'s with different sample sizes? It's easy to see that $\bar{\bar{Y}}$ is unbiased:

$$\begin{aligned} E[\bar{\bar{Y}}] &= E\left[\frac{1}{r}\sum_{i=1}^{r}\bar{Y}_i\right] = \frac{1}{r}\sum_{i=1}^{r} E\left[\bar{Y}_i\right] \\ &= \frac{1}{r}\sum_{i=1}^{r} E\left[\frac{1}{n_i}\sum_{i=1}^{n_i} Y_{ij}\right] = \frac{1}{r}\sum_{i=1}^{r}\frac{1}{n_i}\sum_{i=1}^{n_i} E\left[Y_{ij}\right] \\ &= \frac{1}{r}\sum_{i=1}^{r}\frac{1}{n_i} n_i\mu_Y = \frac{1}{r}\sum_{i=1}^{r}\mu_Y = \mu_Y. \end{aligned} \tag{3}$$

The variance of $\bar{\bar{Y}}$ is similarly easy to derive.

$$\begin{aligned} Var(\bar{\bar{Y}}) &= Var\left(\frac{1}{r}\sum_{i=1}^{r}\bar{Y}_i\right) = \frac{1}{r^2}\sum_{i=1}^{r} Var(\bar{Y}_i) \\ &= \frac{1}{r^2}\sum_{i=1}^{r}\frac{\sigma_Y^2}{n_i} = \frac{\sigma_Y^2}{r^2}\sum_{i=1}^{r}\frac{1}{n_i} \end{aligned} \tag{4}$$

Note that when sample sizes are identical equation (4) yields the familiar minimum variance unbiased estimate (MVUE) result. With equal sample sizes $n_i = N/r \quad \forall i$, and thus

$$Var(\bar{\bar{Y}}) = \frac{\sigma_Y^2}{r^2}\sum_{i=1}^{r}\frac{1}{n_i} = \frac{\sigma_Y^2}{r^2}\sum_{i=1}^{r}\frac{r}{N} = \frac{\sigma_Y^2}{r^2}\frac{r^2}{N} = \frac{\sigma_Y^2}{N}.$$

$Var(\bar{\bar{Y}})$ is minimized when the sample sizes are all equal, and maximized when $r-1$ of the $\bar{Y}_i$'s have a single observation and the one remaining $\bar{Y}_i$ has the remaining $N-(r-1)$ observations.

The preferred approach, which is well known in classical statistics (Kutner et al. 2005) but has been ignored or overlooked in simulation other than by Sánchez and White (2011), is to use a weighted average estimator of the form

$$\bar{\bar{Y}}_w = \sum_{i=1}^{r} w_i \bar{Y}_i \qquad \text{s.t. } \sum_{i=1}^{r} w_i = 1, \quad w_i \geq 0 \quad \forall i \tag{5}$$

and the corresponding unbiased variance estimator is

$$s^2_{\bar{\bar{Y}}} = \left(\frac{\sum w_i}{(\sum w_i)^2 - \sum w_i^2} \right) \sum_{i=1}^{r} w_i (\bar{Y}_i - \bar{\bar{Y}})^2. \tag{6}$$

Using a convex weighting scheme like this results in an unbiased estimator, but the choice of weights affects the variance of the estimate of the sample mean. The set of weights which minimizes the variance of $\bar{\bar{Y}}_w$ is $\{w_i = n_i/N\}$, in which case

$$\begin{aligned} Var(\bar{\bar{Y}}_w) &= Var(\sum_{i=1}^{r} w_i \bar{Y}_i) \\ &= \sum_{i=1}^{r} w_i^2 Var(\bar{Y}_i) \;=\; \sum_{i=1}^{r} \left(\frac{n_i}{N}\right)^2 \frac{\sigma_Y^2}{n_i} \\ &= \frac{\sigma_Y^2}{N^2} \sum_{i=1}^{r} n_i \;=\; \frac{\sigma_Y^2}{N^2} N \;=\; \frac{\sigma_Y^2}{N}. \end{aligned} \tag{7}$$

In other words, with the recommended weights the weighted estimator is MVUE regardless of the varying sample sizes, while the naive estimator is only MVUE when the sample sizes are equal.

3.1 Worst Case Behavior

We can assess the relative impact (RI) of using the naive estimator by looking at the ratio of the variance of $\bar{\bar{Y}}$ to that of $\bar{\bar{Y}}_w$:

$$RI = \left(\frac{\sigma_Y^2}{r^2} \sum_{i=1}^{r} \frac{1}{n_i} \right) / \left(\frac{\sigma_Y^2}{N} \right) = \frac{N}{r^2} \sum_{i=1}^{r} \frac{1}{n_i}. \tag{8}$$

Another way of looking at this is to recall that N is the sum of the n_i's, yielding

$$RI = \left(\frac{1}{r} \sum_{i=1}^{r} n_i \right) \left(\frac{1}{r} \sum_{j=1}^{r} \frac{1}{n_j} \right). \tag{9}$$

Thus the relative impact is simply the average sample size per replication times the average of the inverse sample sizes.

It's easy to confirm that RI is 1 when the n_i's are equal, and increases for any other combination of sample sizes. The maximum variance sampling scenario yields

$$RI_{\max} = \frac{N}{r^2} \left[r - 1 + \frac{1}{N-(r-1)} \right]. \tag{10}$$

For example, if you have ten observations total and five replications, the worst case scenario is that four of the replications produce averages based on a single observation while the remaining replicate produces an average of six observations. In that case the naive average of the averages would have $\frac{5}{3}$ the variance of the weighted estimator.

3.2 An Interesting Alternative Measure

The maximum variance scenario is unlikely to occur in practice, so arguably doesn't give much insight other than by bounding RI. At this point we have no idea of whether that's a loose bound or a tight one. Consider the following alternative measure. Suppose the number of replications is even and that half of the replicates have the same sample size n_{low} while the other half have sample size n_{hi}. The total sample size is then $N = (r/2)(n_{low} + n_{hi})$. Substituting this into equation (8), we get

$$\begin{aligned} RI_I &= \left(\frac{1}{r^2}\right)\left(\frac{r}{2}\right)(n_{low}+n_{hi})\left(\sum_{i=1}^{r/2}\frac{1}{n_{low}}+\sum_{j=1}^{r/2}\frac{1}{n_{hi}}\right) \\ &= \left(\frac{1}{2r}\right)(n_{low}+n_{hi})\left(\frac{r}{2}\left[\frac{1}{n_{low}}+\frac{1}{n_{hi}}\right]\right) \\ &= \left(\frac{1}{4}\right)(n_{low}+n_{hi})\left(\frac{1}{n_{low}}+\frac{1}{n_{hi}}\right) \\ &= \left(\frac{1}{4}\right)\left(2+\frac{n_{hi}}{n_{low}}+\frac{n_{low}}{n_{hi}}\right). \end{aligned} \tag{11}$$

Note that this measure is invariant in both N and r, hence the subscript I. If we describe the observed sample sizes as a set $\{n_1, n_2, \ldots, n_r\}$, this says that experiments with $\{1,4\}$, $\{25,100\}$, and $\{1,1,1,1,1,4,4,4,4,4\}$ all have the same sample-size ratio $n_{hi}/n_{low} = 4$, and therefore all have the same relative impact $RI = 1.5626$. The estimated variance of the mean will be more than 50% larger when using $\bar{\bar{Y}}$ rather than $\bar{\bar{Y}}_w$.

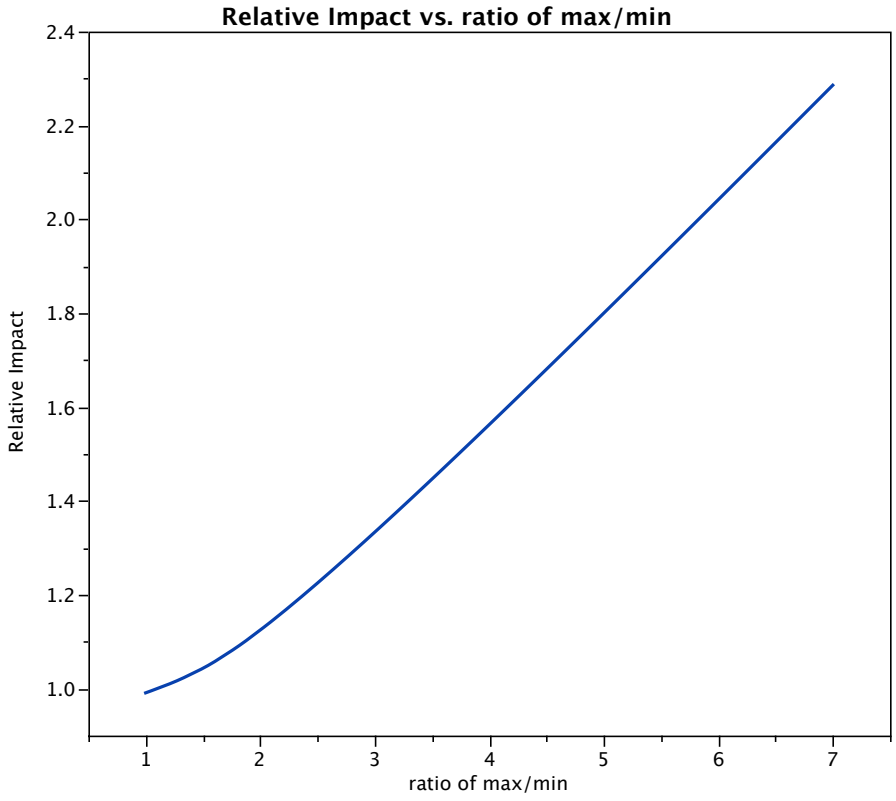

Figure 1: Invariant relative impact vs. the ratio of maximum sample size to minimum sample size.

Figure 1 shows the variance inflation on the vertical axis as a function of the sample-size ratio along the horizontal axis. As an example, a 5:1 sampling ratio inflates variance by a factor of 1.8 when the naive estimator is used. Note that the relative impact is non-linear for sample-size ratios below 2 but becomes very nearly linear for higher ratios.

4 AN EXAMPLE: TURNING LEMONS INTO LEMONADE

Consider a system that processes items serially, one at a time. Each item is inspected after processing. If the item fails inspection, it is reprocessed and inspected once again. The reprocessing/inspection cycle continues indefinitely until the item finally passes inspection. Items failing inspection are "lemons", and the probability that a lemon passes subsequent inspections after any reprocessing/inspection cycle is far less than the initial probability that the item is a lemon.

For this example, the total processing-and-inspection time on any cycle is random variable P distributed $TRI(25,30,35)$ minutes with $E[P] = 30$. The probability that an item is not a lemon (and passes on its first inspection) is $p_1 = 0.95$. The probability that a lemon passes inspection after any reprocessing/inspection cycle is $p_r = 0.10$. The simulation terminates after at least 480 minutes have expired and any item still in process passes inspection. That is, the system will work overtime beyond a normal 8-hour day if any item is still in process. The system begins with a new item immediately available and ready for processing. That is, the system regenerates after the terminating condition is achieved and there is no initialization bias.

The objective is to estimate the average processing time for all items. We can compute this average analytically. If the item is a lemon, the number of reprocessing/inspection cycles incurred before the final cycle which results in the item passing inspection has a $GEOM(p_r)$ distribution. The expected processing time is therefore

$$\begin{aligned} E[T] &= E[P]\left(p_1 + (1-p_1)\left(\frac{1}{p_r}\right)\right) \\ &= 30\left(0.95 + 0.05\left(\frac{1}{0.1}\right)\right) \\ &= 30(1.45) = 43.5. \end{aligned} \tag{12}$$

We ran 100 simulation replications for this system and divided the results into 5 sequential experiments of 20 replications each. The resulting 95% confidence intervals for each 20-replicate experiment, as well for the single combined 100-replicate experiment, are shown in Figure 2. These intervals are computed using both the traditional approach (displayed as lighter-colored bars to the left) and the optimal weighting scheme (displayed as darker bars to the right). The true mean derived in equation 12 is plotted as a horizontal red line. Note that while all of the point estimates are greater than the true mean—an artifact of the terminating condition which we have not corrected—all but one of the intervals cover. The exception is the combined experiment using the traditional approach. This is noteworthy because without prior knowledge of the correct answer most analysts would probably consider the combined estimator to be more reliable because of its larger sample size. Note also that *the weighted estimators provided both greater accuracy and greater precision in all cases.*

Figure 3 is a scatter plot of the number of observations versus the estimated mean for each of the 100 replications. Note that the weights applied are proportional to the number of observations for each replication. The power trendline illustrates that larger numbers of observations are associated with smaller estimated average processing times. The beneficial effect of the weighting scheme is clear—larger weights are given to the more common outcomes where large reprocessing delays are unusual, since lemons represent only 5% of all items in the general population.

Note that this example was not difficult to construct. It is representative of regenerative processes subject to conditions that tend to correlate the magnitude of an aggregate output measure with the number of observations obtained during any regenerative cycle. In this example, the events are "disruptive" and

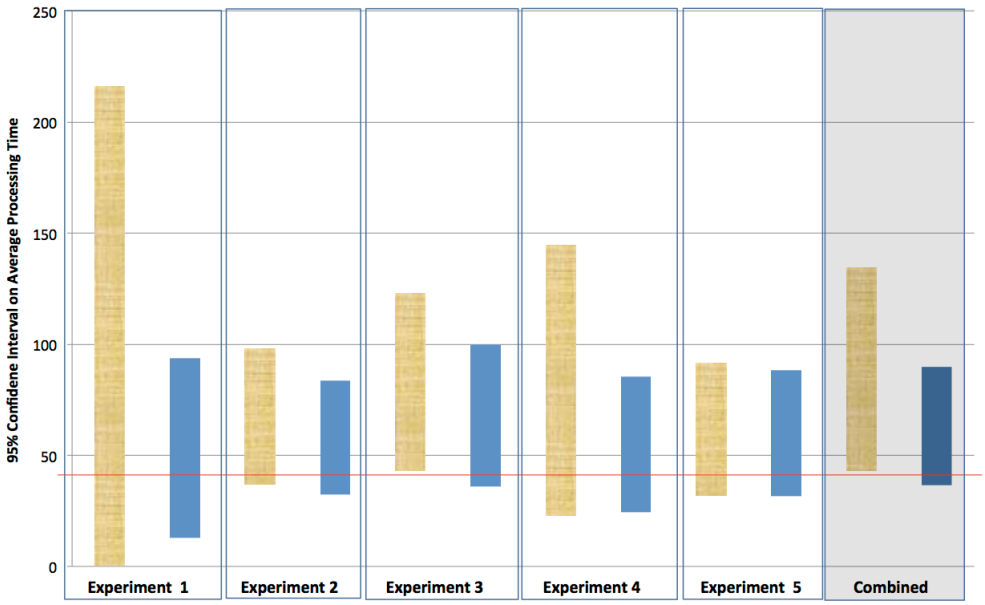

Figure 2: 95% confidence intervals on the mean processing time computed using both the traditional approach (lighter bars to the left) and the recommended weighting scheme (darker bars to the right) for the 20-replication experiments and the combined 100-replication experiment.

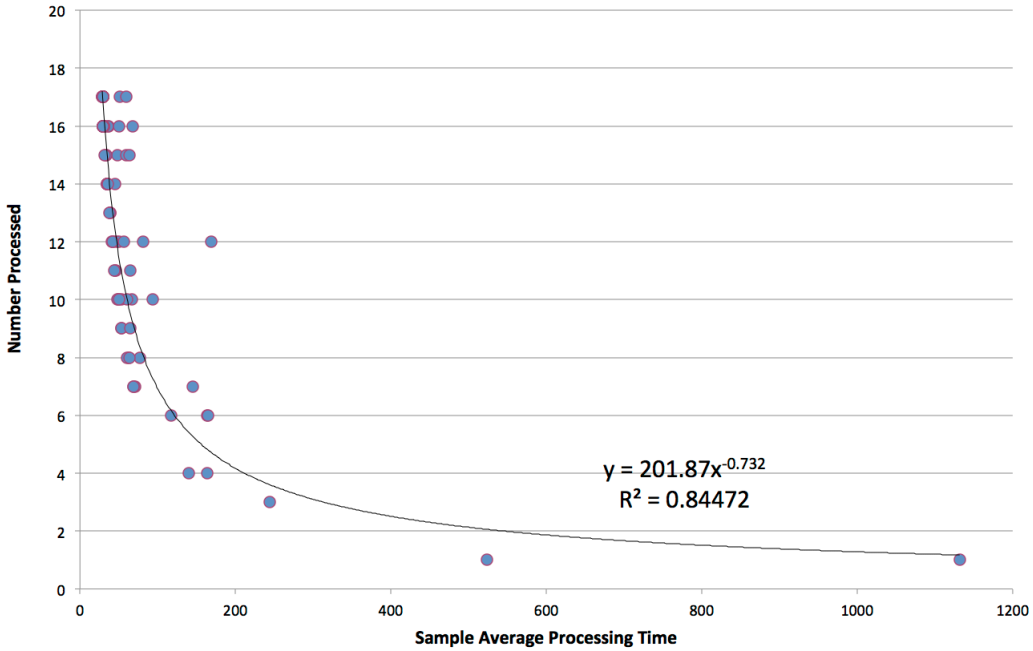

Figure 3: Scatter plot illustrating negative correlation between the number of items completed and the estimated mean processing time for each of 100 replications.

the correlation is negative, as might be the case in systems with random failures or vacations. One can also think of events that are "benign" and induce a positive correlation, such as unusually long stretches of good weather and/or instances of unusually low absenteeism on crews for construction projects.

5 CONCLUSIONS

Using the naive $\bar{\bar{Y}}$ estimator rather than $\bar{\bar{Y}}_w$ will often make little or no difference. This is the case when the terminating simulation produces 1) single-observation measures as its output; 2) aggregate measures where the terminating rule guarantees identical sample sizes; or 3) the sample sizes can vary from run to run but are strongly consistent, i.e., the ratio of max to min sample size is relatively close to one. In cases 1) and 2), the naive estimator and the weighted estimator are mathematically identical. In case 3) the two estimators should produce results that are very close to each other, but to the extent that the estimates differ the weighted estimator is the better one. When none of the three cases applies, the weighted estimator can significantly outperform the naive estimator. Given that the weighted estimator is simple to compute, we recommend that it should always be used.

ACKNOWLEDGMENTS

The authors would like to thank Averill Law for several interesting discussions about this topic. We would also like to thank the referees for their valuable feedback.

REFERENCES

Banks, J., J. S. Carson, B. L. Nelson, and D. M. Nicol. 2000. *Discrete-Event System Simulation*. 3rd ed. Upper Saddle River, New Jersey: Prentice-Hall, Inc.

Bratley, P., B. L. Fox, and L. Schrage. 1983. *A Guide To Simulation*. New York: Springer-Verlag.

Hoover, S. V., and R. F. Perry. 1989. *SIMULATION: A Problem-Solving Approach*. Addison-Wesley Publishing Company, Inc.

Kutner, M. H., C. J. Nachtsheim, J. Neter, and W. Li. 2005. *Applied Linear Statistical Models*. 5 ed. McGraw-Hill, Inc.

Law, A. M., and W. D. Kelton. 2000. *Simulation Modeling & Analysis*. 3rd ed. New York: McGraw-Hill, Inc.

Sánchez, P. J., and K. P. White, Jr.. 2011. "Interval Estimation Using Replication/Deletion and MSER truncation". In *Proceedings of the Winter Simulation Conference*, edited by B. Johansson, S. Jain, R. R. Creasey, J. Himmelspach, K. P. White, Jr., and M. Fu, 488–494. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

AUTHOR BIOGRAPHIES

PAUL J. SÁNCHEZ is a faculty member in the Operations Research Department at the Naval Postgraduate School. His research focuses on the intersection between computer modeling and statistics. In his spare time he enjoys reading science fiction and mysteries. His email address is pjsanche@nps.edu and you can check out some of his work at <http://harvest.nps.edu>.

K. PRESTON WHITE, JR. is Professor of Systems Engineering at the University of Virginia. He is a past member of the WSC Board of Directors and was General Chairman for WSC2011. He received the B.S.E., M.S., and Ph.D. degrees from Duke University. He has held faculty appointments at Polytechnic University and Carnegie-Mellon University and served as Distinguished Visiting Professor at Newport News Shipbuilding, at SEMATECH, and at the Naval Postgraduate School. He is a member of INFORMS and a senior member of IEEE and IIE. He sits on the Advisory Board of VMASC. His email address is kpwhite@virginia.edu.